# PawlakDominik_HW1

2022-03-30

```
library(DALEX)
```

```
## Welcome to DALEX (version: 2.4.0).
## Find examples and detailed introduction at: http://ema.drwhy.ai/
## Additional features will be available after installation of: ggpubr.
## Use 'install_dependencies()' to get all suggested dependencies
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
set.seed(2137)

data <- read.csv("insurance.csv")
head(data)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Now, let's split the data into training and test datasets.

```
index <- createDataPartition(apartments$m2.price, p = 0.5, list = FALSE)

train <- data[index,]
test  <- data[-index,]

x_train <- train[,-c(7)]
y_train <- train[, 7]

x_test <- test[,-c(7)]
y_test <- test[, 7]
```

After splitting the data, we can train the model.

```
library(ranger)

forest <- ranger(charges~., data=train)
y_pred <- predict(forest, x_test)
print(y_pred$predictions[50])
```

```
## [1] 21943.25
```

```r
print(y_test[50])
```

```
## [1] 15820.7
```

Let's create explainer, then BreakDown Composition for this observation.

```r
explainer_rf <- DALEX::explain(forest,
                               data = x_test,
                               y = y_test)
```

```
## Preparation of a new explainer is initiated
##   -> model label       :  ranger  (  default  )
##   -> data              :  837  rows  6  cols
##   -> target variable   :  837  values
##   -> predict function  :  yhat.ranger  will be used (  default  )
##   -> predicted values  :  No value for predict function target column. (  default  )
##   -> model_info        :  package ranger , ver. 0.13.1 , task regression (  default  )
##   -> predicted values  :  numerical, min =  2278.106 , mean =  13280.87 , max =  44886.47
##   -> residual function :  difference between y and yhat (  default  )
##   -> residuals         :  numerical, min =  -7738.32 , mean =  79.62459 , max =  29440.24
##   A new explainer has been created!
```

```r
bd_pr <- predict_parts(explainer = explainer_rf,
                       new_observation = x_test[50,],
                       type = "break_down")
bd_pr
```

```
##                              contribution
## ranger: intercept               13280.866
## ranger: smoker = yes            -4585.926
## ranger: bmi = 19.3               -741.087
## ranger: age = 38                -1352.486
## ranger: children = 0             -834.768
## ranger: region = southwest        408.993
## ranger: sex = male                247.088
## ranger: prediction               6422.682
```

Now let's check Shapley values

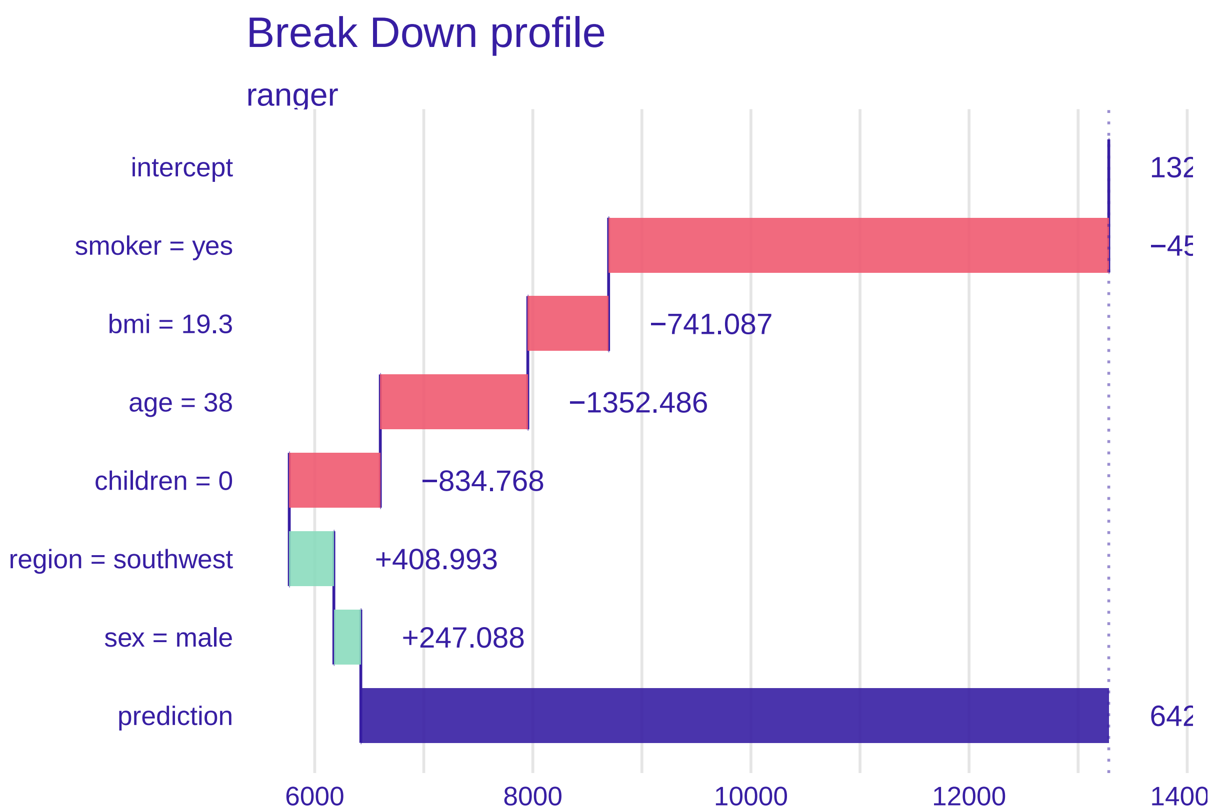
```r
shap_pr <- predict_parts(explainer = explainer_rf,
                         new_observation = x_test[50,],
                         type = "shap")
shap_pr
```

```
##                                    min         q1     median       mean
## ranger: age = 38            -2068.7206 -1820.2292 -1606.9697 -1642.7651
## ranger: bmi = 19.3          -2750.1462 -2241.2093  -930.5814 -1332.4012
## ranger: children = 0         -904.6873  -710.6610  -672.8803  -668.4678
## ranger: region = southwest    102.0011   377.1730   597.0170   598.7342
## ranger: sex = male            194.2034   247.0881   283.9365   302.5775
## ranger: smoker = yes        -4978.1031 -4738.5023 -4513.8531 -4115.8624
##                                     q3        max
## ranger: age = 38            -1460.8523 -1368.4610
## ranger: bmi = 19.3           -832.6867  -128.2763
## ranger: children = 0         -614.3792  -403.1087
## ranger: region = southwest    719.6988  1004.7095
## ranger: sex = male            326.0912   541.0396
```

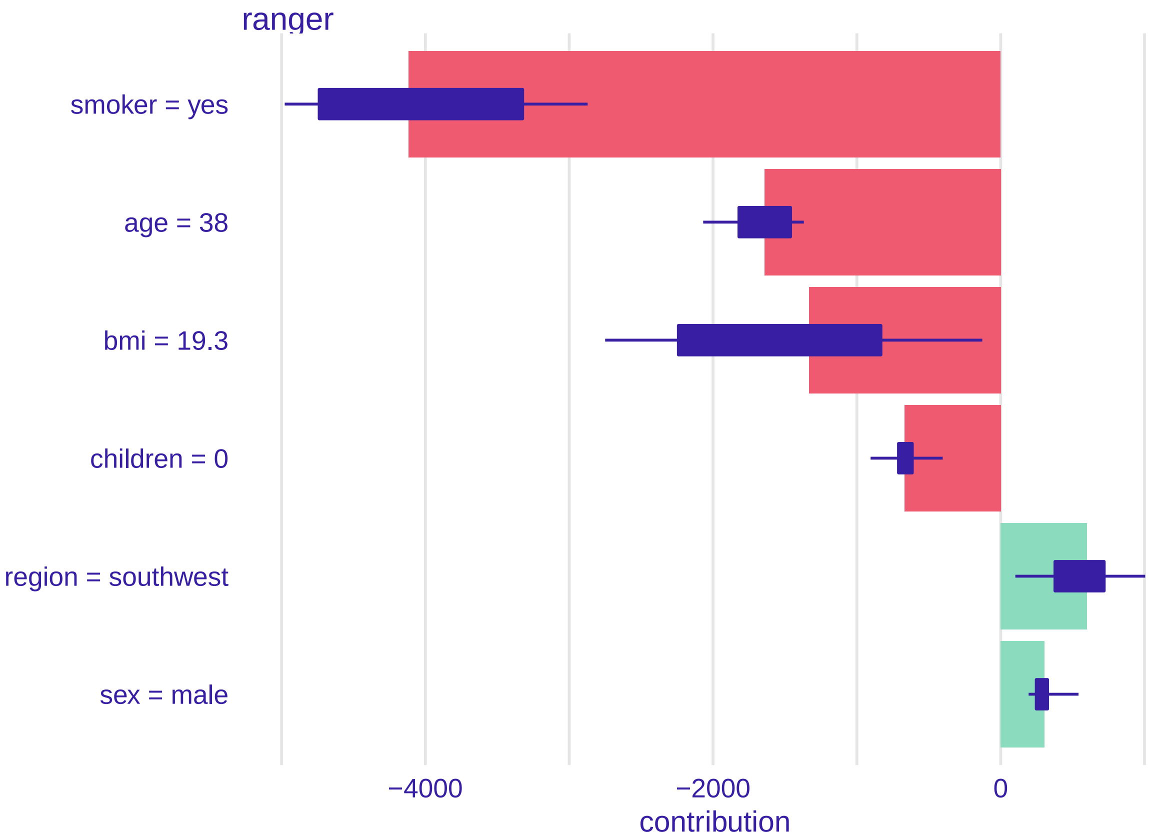
```
## ranger: smoker = yes        -3323.6158 -2871.8476
```

Let's plot and compare both charts

```
plot(bd_pr)
```




```
plot(shap_pr)
```




Both plots suggest that variable smoker, with “yes” value has the biggest impact on the prediction and decreases the result. The variable “age” also decreases the prediction. Both plots suggest that ‘sex’ variable doesn’t have big influence on the result. According to Break Down decomposition the region variable increases the prediction, whereas the according to the shapley values, this variable decreases it.

Now, let’s find a female who doesn’t smoke and check the results for that person.

```
observation2 <- test[(test$sex=="female" & test$smoker=="no" & test$age >= 64),]
observation2 <- observation2[1,]
observation2
```

```
##      age    sex   bmi children smoker    region  charges
## 200   64 female 39.33        0     no northeast 14901.52
```

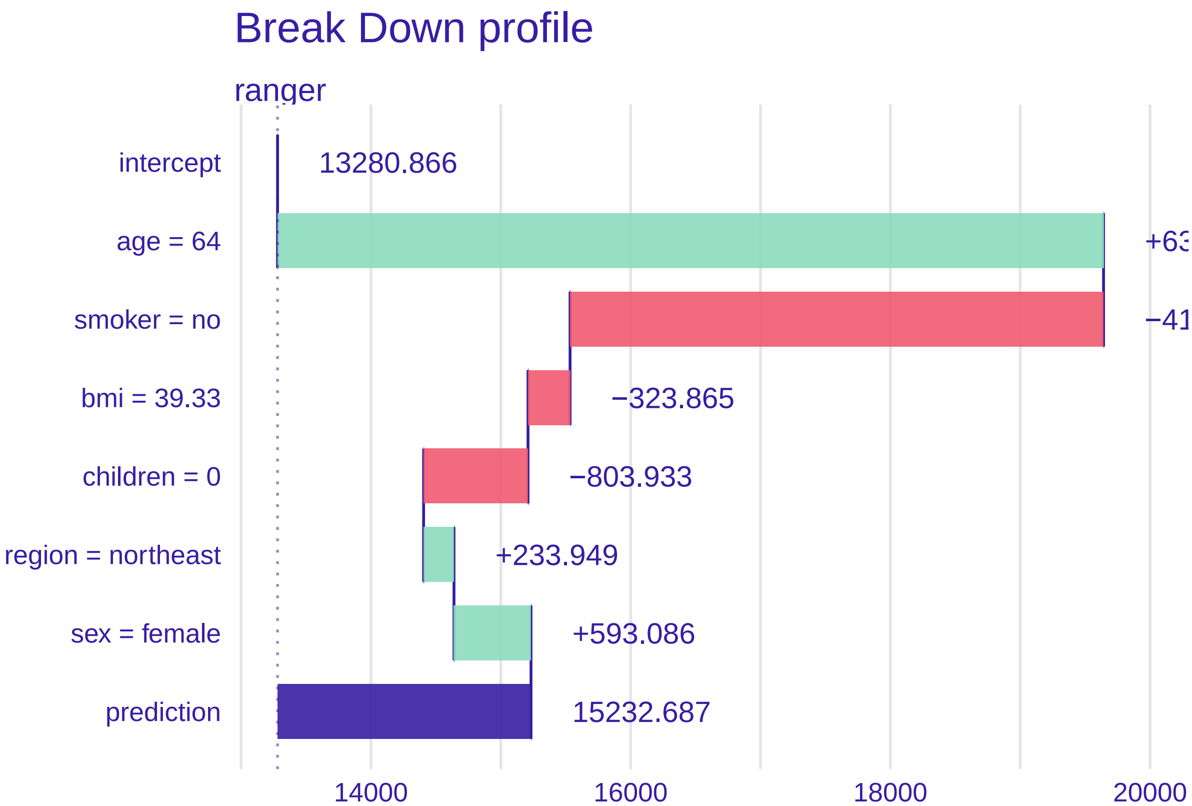
Now let’s repeat steps for this observation.

```
bd_pr <- predict_parts(explainer = explainer_rf,
                       new_observation = observation2,
                       type = "break_down")
bd_pr
```

```
##                             contribution
## ranger: intercept              13280.866
## ranger: age = 64                6360.833
## ranger: smoker = no            -4108.250
## ranger: bmi = 39.33             -323.865
## ranger: children = 0            -803.933
## ranger: region = northeast       233.949
## ranger: sex = female             593.086
## ranger: prediction             15232.687
```

And plot it:

```
plot(bd_pr)
```




Now

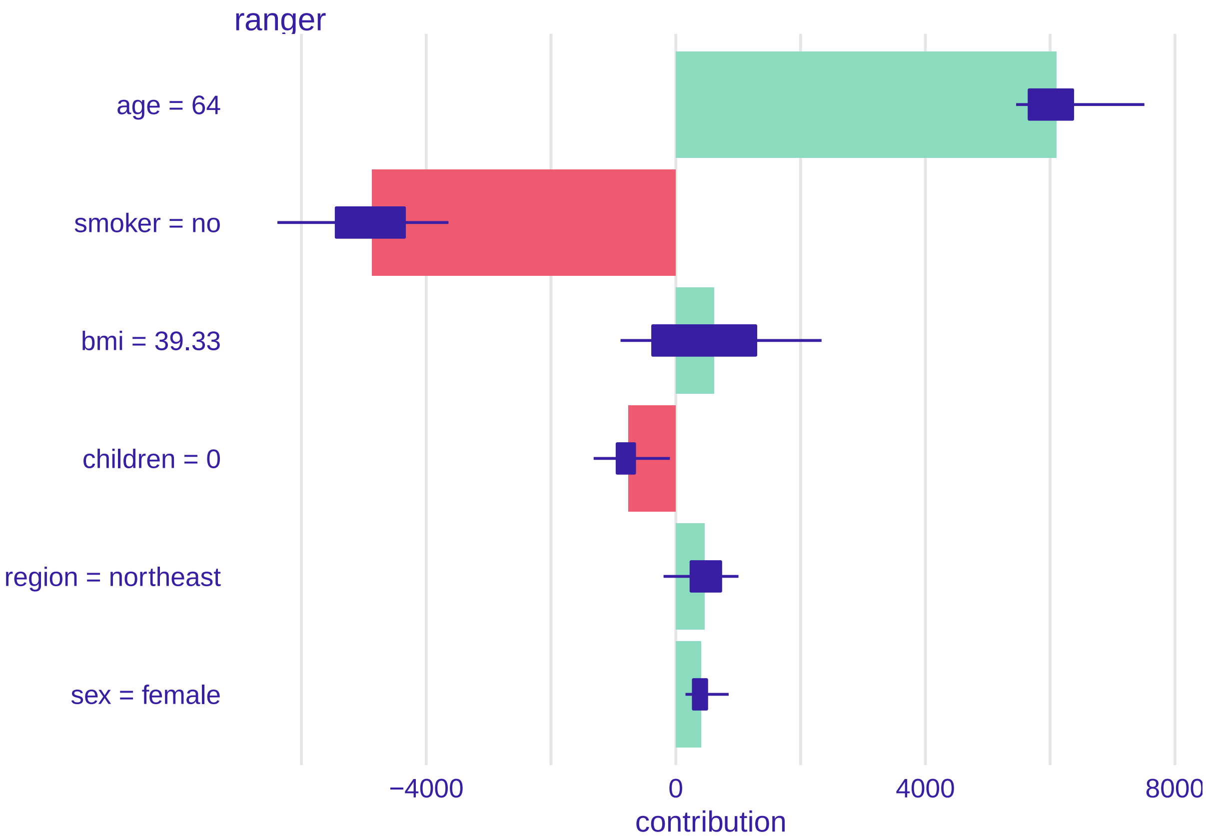
let's check Shapley values

```
shap_pr <- predict_parts(explainer = explainer_rf,
                         new_observation = observation2,
                         type = "shap")
shap_pr
```

```
##                                     min         q1     median       mean
## ranger: age = 64              5454.4151  5663.0079  5993.1766  6100.9647
## ranger: bmi = 39.33           -885.3304  -370.2400   478.4041   615.9200
## ranger: children = 0         -1316.7621  -940.5594  -687.4544  -756.2491
## ranger: region = northeast    -196.3346   244.2341   528.5900   460.1631
## ranger: sex = female           155.1834   281.1589   350.9949   400.7234
## ranger: smoker = no          -6384.9140 -5442.6144 -4629.4526 -4869.7019
##                                      q3         max
## ranger: age = 64              6360.8331  7510.26653
## ranger: bmi = 39.33           1281.7362  2336.22239
## ranger: children = 0          -660.7291   -94.58414
## ranger: region = northeast     719.6988  1004.70950
## ranger: sex = female           494.6683   847.07052
## ranger: smoker = no          -4349.0123 -3643.55826
```

And plot it:

```
plot(shap_pr)
```




Conclusions: In the first observation, both plots present “smoker” (yes) as the most significant variable, that decreases the predicted result. On contrary, in the second observation, “age” variable turns out to be the most significant and it increases the predictions. What’s more interesting and surprising the “smoker” (no) decreases the predictions. In both scenarios “children” and “sex” variables seem not to have big impact on the result. In the first case, “region” variable has also big impact, wheras in the second observation, it hardly affects the predictions.